The Report Committee for Juan Carlos Santos
Certifies that this is the approved version of the following report:

The Implementation of Phylogenetic Structural Equation Modeling for Biological Data from Variance-Covariance Matrices, Phylogenies, and Comparative Analyses

APPROVED BY
SUPERVISING COMMITTEE:

Supervisor:

Claus O. Wilke

James Bryant

The Implementation of Phylogenetic Structural Equation Modeling for Biological Data from Variance-Covariance Matrices, Phylogenies, and Comparative Analyses

by

Juan Carlos Santos, B.A., Ph.D.

Report

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science in Statistics

The University of Texas at Austin

December, 2009

Dedication

To Natalia and Ignacio the best “Black Swan Events” of my life

To my parents, Ernesto and Sara, for let me follow the unpredictable

Acknowledgements

This report is the result of unexpected events related to my doctoral thesis and the data that was generated. I thank Claus Wilke for his insights during the writing process and his encouragement during the last part of this graduate program. I thank James Bryant for let me be his TA for Biostatistics and his helpful comments to improve this report. I thank David Cannatella, my PhD advisor, for his support and understanding during my wanderings in statistics. I thank Natasha Beretvas and Tiffany Whittaker for their courses that taught me the wonders of factor analysis and SEM, their lectures and exercises help me in the development of this report. I thank Bob Penman for his help in correcting and formatting this report.

Finally, I thank, to my parents, Ernesto and Sara, to my brothers, Mauricio and Fabian for their support and love. I thank Natalia and Ignacio for their support, encouragement, patience, and love during this hectic time of finishing a PhD, Master, and moving out of Texas.

December, 2009

Abstract

The Implementation of Phylogenetic Structural Equation Modeling for Biological Data from Variance-Covariance Matrices, Phylogenies, and Comparative Analyses

Juan Carlos Santos, M.S.Stat.

The University of Texas at Austin, 2009

Supervisor: Claus O. Wilke

One statistical approach with a long history in the social sciences is a multivariate method called Structural Equation Modeling (SEM). The development of SEM followed the evolution of factor and path analyses, multiple regression analysis, MANOVA, and MANCOVA. One of the key innovations of factor analysis and SEM is that they group a set of multivariate statistical approaches that condense variability among a set of variables in fewer latent (unobserved) factors. Most biological systems are multivariate, which are not easily dissected into their component parts. However, most biologists use only univariate statistical methods, which have definitive limitations in accounting for more than a few variables simultaneously. Therefore, the implementation of

methodologies like SEM into biological research is necessary. However, SEM cannot be applied directly to most biological datasets or generalized across species because of the hierarchical pattern of evolutionary history (i.e., phylogenetic non-independence or signal). This report includes the theoretical grounds for the development of phylogenetic SEM in preparation of the development of utilitarian algorithms. I have divided this report in six parts: (1) a brief introduction to factor analysis and SEM from a historical perspective and a brief description of their utility; (2) a summary of the implications of using biological data and their underlying hierarchical structure due to shared ancestry or phylogeny; (3) a summary of the two most common comparative methods that incorporate phylogenies in their implementation (i.e., phylogenetic independent contrasts, and phylogenetic generalized least squares); (4) I describe how independent contrasts and correlation matrices from both comparative methods can be used to estimate a phylogenetically corrected variance–covariance matrices; (5) I describe how to perform a exploratory factor analysis, specifically phylogenetic principal component analysis, with the corrected variance–covariance matrices; and (6) I describe the development of the phylogenetic confirmatory factor analysis and phylogenetic SEM. I hope that this report encourages other researchers to develop adequate multivariate analyses that incorporate the evolutionary principles of shared ancestry and phylogeny in their estimations.

Table of Contents

List of Figures

Chapter 1: A Primer in Structural Equation Modeling and its Implementation in Evolutionary Biology

The development of structural equation modeling followed the evolution of factor and path analyses. Factor analysis groups a set of multivariate statistical approaches that condense variability among a set of variables in fewer latent (unobserved) factors. Charles Sperman developed factor analysis in the early 1900's after the establishment of the multivariate regression analysis by Karl Pearson at the end of the XIX century (Brown 2006). The emergence of factor analysis followed the development of psychometrics in an attempt to determine underlying patterns in otherwise unmanageable matrices of correlated data (Mulaik 2009). Most methods in factor analysis try to summarize correlation-covariance matrices using linear combinations of factors and remaining unexplained variability (error) (Brown 2006). The information condensation is used to reveal correlations and pathways among observed variables, that otherwise will not be apparent from raw data.

Among biologist the most methodology used to summarize variability is principal component analysis (PCA) (Grace 2006). Factor analysis is related to PCA but differs in the inference about the determinants of the variance-covariance matrices among the observed variables. In PCA, we try to maximize the variance explained among set observed variables by uncorrelated components that rotate in variable space. In factor analysis, we try to determine a set of latent common factors that explain the variability among observed variables. Therefore, the evolution of factor analysis in a method incorporates unseen common factors and observed variables evolved to incorporate path analysis.

Sewall Wright is regarded among the founders of theoretical population genetics and the initiators of the modern evolutionary synthesis of evolution and genetics (Huxley et al. 1942). However, Sewall Wright also had a profound influence in development of multivariate statistics including path analysis (Rao et al. 2000). Path analysis was original developed on the 1920's to analyze inbreeding and systems of mating (Wright 1918, 1920, 1921, 1923, 1925, 1934). However, the was not fully applied until maximum likelihood estimation and likelihood ratio test for hypothesis test was introduced in the late 1960's in Social Sciences (Li 1975). The implementation of path analysis in ecology only started in the 1990's following the implementation of software for path analysis (e.g., AMOS, EQS, and LISREL) (Grace 2006). The development of path analysis, factor analysis, and the maximum likelihood estimation has developed to give rise to structural equation modeling (SEM) in the 1990's to the present (Kline 2005).

Path analysis can described as structural linear multiple regression with two main applications: (1) to model the relationships among a set of observed and correlated variables, and (2) to evaluate the importance of observed measurements on describing latent (unobserved) variables (Li 1975). Two key feature introduced by Sewall Wright (Wright 1921) was introduction of model causation analysis among correlated variables and the idea of latent variables (Figure 1). Path analysis provides a mean to test causal relationships by testing hypothesized (theoretical) models of variable relationships against observed data (Li 1975). It is important to emphasize that Sewall Wright in developing path analysis did not attempted to discover causal relationships, but if these are hypothesized they can be tested using path analysis (Wright 1923, 1983). The application of path analysis is more in terms to explain correlation and not causation of variables, and many models may be consistent with an empirical dataset. Therefore, path analysis provides a statistical approach to determine the relative importance of alternative

models of variables relationships in multivariate scheme (Li 1975). The current terminology refers to path analysis as the modeling single-indicator variables without latent variables (Kline 2005).

The implementation of path analysis starts with the researcher proposing a model of the variable relationships (Brown 2006). The investigator usually has some previous knowledge of variables relationships that allow him to propose a causal hypothesis (i.e., theoretical grounds or independent evidence of variable relationship). The hypothesis can be draw as a diagram with causal relationships of variables including independent, intermediate, dependent, and error (Figure 2). Independent variables are at the top in a chain of causation path (i.e., are exogenous) and two or more independent variables can be associated by correlation. Intermediate variables are connecting and dependent (i.e., endogenous and exogenous) steps in a chain of causation. Dependent variables are always influenced (i.e., endogenous) by other variables including independent, intermediates, or other dependents. Error terms represent the remaining variance unexplained by the proposed model that reflects the effect on endogenous variables by unmeasured exogenous variables outside model and measurement error. Therefore, the model can be stratified with several levels of causal connections (i.e., with intermediate variables in the path of causation).

The variable relationship diagram (path diagram) and the significance of the path coefficients connecting variables are used to perform the inference in path analysis (Figure 1 and 2). The graphical representation of the paths is arrows linking variables. Single arrows represent causation between exogenous (independent, intermediate dependent, and error variables) and endogenous (dependent variables). Double-headed arrows indicate correlation between pairs of exogenous variables (independent, dependent, or errors). From the path diagram, correlations among variables can be

derived from the path coefficients using a set of rules (Wright 1921; Li 1975). First, the correlation between two variables can be determined by the sum of the products of the path coefficients that join both variables. For example, the correlation between two variables joined by a single path (direct) and by two paths through an intermediate (indirect path) will be the sum of the products of both pathways (i.e., direct and indirect). Second, the significance of a causal influence of an independent over a dependent variable is measured by the square of the path coefficient (i.e., determination coefficient). Alternatively, the interpretation of the path coefficients can be given in terms of multiple regression analysis (Grace 2006). The path coefficient is a standardized regression coefficient reflecting the magnitude of the direct effect of an exogenous (i.e., independent or intermediate variable) on an endogenous (i.e., intermediate or dependent variable) in the path model. Thus, partial regression coefficients between a dependent and two or more exogenous variables can also be estimated using path analysis. Therefore, the path coefficients can also measure the magnitude of the effect of variable on another after controlling for other related variables.

The interpretation of the path coefficients allows several inferences about relationships between independent and depend variables. First, the path multiplication rule allows the interpretation of the effect of an independent variable on a dependent through intermediate variables. For example, if A is exogenous variable that connects to endogenous variable C through an intermediate endogenous variable B (i.e., A $\rightarrow$ B $\rightarrow$ C), the effect of A on C should be equal to the product of the coefficient paths $b_{AB} * b_{BC}$. Second, total causal effect decomposition of an endogenous can be inferred from path analysis by adding direct and indirect effects. Direct effects represent direct paths between exogenous and endogenous variables, which are usually evidenced by pairwise correlation analysis. Indirect effects represent a significant conceptual development by

providing an explanation of the variability on endogenous variables from intervening exogenous variables in the path.

Path analysis is extremely flexible in terms model construction, but it has some assumptions and requirements in order to be applicable (Grace 2006). First, the relationships among variables have to linear, but variables can be transformed to meet the assumption. Second, error term of any endogenous variable is uncorrelated with any other endogenous variable in the model. Third, variables must show some level of correlation but multicollinearity must be low. Fourth, model should over-identified or at least just identified in order to be estimated, that is that to estimate the path coefficients an equal or more structural equations have to be provided. Fifth, a proper correlation or variance-covariance is required as input. Sixth, large sample sizes and completeness might be necessary to be able get an adequate estimation of the path coefficients and significance.

The evolution of path analysis into structural equation modeling (SEM) started by incorporating models with latent variables measured by multiple observable variables. SEM was extended incorporating model fitting indices and respecification patterns (paths) of variable relationships, after determining their goodness of fit and fitting indices (Kline 2005). SEM was developed to be flexible in model construction by incorporating correlations among all types of variables (independent, dependent, and errors), latent variables, multilevel interaction (latent, intermediate, and dependent variables), and latent with multiple indicator (observable) variables (i.e., confirmatory factor analysis). In general, SEM is considered to be an extension of general linear model (GLM) and incorporates concepts of factor analysis, multiple regression, and path analysis. Other advantages of SEM include graphical visualization of models; model test with all coefficients simultaneously; the ability to model error terms; and to incorporate temporal autocorrelated errors (e.g., approximation to time series). However, as in path analysis,

SEM cannot be use to infer causation and the interpretation of model is based on theoretical grounds underlying the research.

SEM applications include data description (exploratory) and model fit (confirmatory). Most researchers regard the confirmatory approach of SEM as the most suited for its features and it can use in three approaches (Brown 2006). First, strictly confirmatory or single model goodness-of-fit of the proposed structural path to the pattern of variances and covariances of the observed data. Second, model testing of alternative structural models to the observed data with series model comparisons by LTR and fit indices. Third, model development or a combination of exploratory and confirmatory SEM that includes proposing an initial structural model, rectification of the model, and retesting based on changes suggested by SEM modification indexes. Usually all SEM approaches might require cross-validation of the model developed using an independent validation sample (Kline 2005).

The methodological approach to SEM is divided in two main processes: (1) measurement model validation, and (2) structural model fitting (Brown 2006) (Figure 2). The first process accomplished by proposing a model of variable relationships including latent and indicator variables. Latent variables are hypothesized factors that exist and can only be measured through at least ≥ 2 observable indicator variables (Brown 2006). The model is validated if the indicators measure the corresponding latent variables through common factor analysis or principal axis factoring to the observed data. The last process is accomplished by path analysis by comparing two or more alternative models (one of which may be the null model) in terms of model fitness. Therefore, the best model is the one that best predicts the observed covariances in the data with less number of parameters estimated. The implementation of SEM approach will be detailed in the last chapter of this report.

SEM and path analysis has similar assumptions and requirements (Grace 2006). First, SEM assumes linear relationships the relationships among latent and indicator variables, and between latent variables. Second, multivariate normal distribution of indicator variables is required for MLE of the structural model. Third, latent variables should have > 2 indicator (observable) variables that reduces the measurement error of the structural model. Fourth, indicators should have some level of correlation but multicollinearity must be low. Fifth, model should over-identified or at least just identified in order to be estimated, that is that the number of many parameters to be estimated should be the same or less than the elements in the covariance matrix. Sixth, a proper correlation or variance-covariance is required as input. The original sample can have correlated values and known in order to be incorporated in the model. Finally, large sample sizes might be necessary to be able get an adequate estimation of the structural model and different dataset can be used to cross-validate a structural model.

The implementation of SEM in biology might be straight-forward, but some adjustment and correction to the original data might be required (Brown 2006). First, most biological data is non-independent as species or individuals share different degrees of phylogenetic relationships. Second, the variance-covariance matrix is required as input for SEM in order to be applicable. Third, sample sizes in biological systems might be limited and some penalizations on the SEM fit indices might need to be adjusted. In this report, I will address some of the available methods to correct for phylogenetic signal and estimation of corrected variance-covariance matrix (chapters 2 – 4), and then implementation of phylogenetic factor analysis and SEM will be presented (chapters 5 – 6).

Chapter 2: The Problem of Phylogenetic Non–independence in Comparative Analyses

All organisms descent of at least one common ancestor and with whom they share a joint evolutionary history. Life forms conform lineages of descendents that have an underlying non–independent hierarchical structure based on a temporal patter of divergence from a common ancestor (Mayr 1982). In this context, phylogenetic signal is defined as the tendency of phenotypic resemblance among organisms inherited by their ancestors (Blomberg et al. 2003). Alternatively, the process of natural selection will produce apparent similarities due to convergence and not by shared ancestry (Brooks 1996). Therefore, in comparative evolutionary analysis is necessary to incorporate phylogentic history to differentiate simplesiomorphy (shared traits by descent) from convergence (adaptive or stochastic) (Harvey and Pagel 1991).

Phylogenetic signal has practical consequences on the characterization and inferences about trait evolution among taxa (Felsenstein 2004). First, phenotypes of species are expected to be more similar to values of close relatives. Thus, we can make predictions about species characteristics based only on the phenotypes of closely related taxa (Garland and Ives 2000). Moreover, branch lengths of a phylogeny are integrated as a priory expectation of phenotypic change, which can be incorporated in the construction of models of character evolution (Garland et al. 2004). Second, comparative analysis of phenotypic traits cannot be based on standard statistics as these methods assume that data–points are independent and they share common variance. The recognition of this has allowed the development of methodologies adequate the hierarchical structure evolutionary history and weight the level dependency based on a phylogenetic relationship (Rohlf 2001). Therefore, models of character evolution including phylogenetic information are robust against the bias of shared ancestry by simultaneously

reduce Type I error caused by non–independence (Blomberg et al. 2003). However, some natural processes that affect phylogenetic signal are difficult to model including hybridization, horizontal gene transfer, character displacement, homoplasy, and long periods of stabilizing selection (Wake 1991; Losos 2000; Martins et al. 2002). As a consequence, more complex phylogenetic comparative methods are being developed (e.g., Hansen et al. 2008).

Phylogenetic comparative methods have been important in expanding our inferences about evolutionary processes (Harvey and Pagel 1991). Most researchers use comparative analyses to infer adaptation or response to natural selection (Garland and Adolph 1991; Garland et al. 2004) (Figure 3). Other alternative implementation of the comparative methods include effects of sexual selection (adaptive, and non–adaptive) (Cox et al. 2003), comparisons of rates of phenotypic evolution (Garland and Ives 2000), ecological trade–offs under phylogenetic perspective (Clobert et al. 1998); and phenotypic integration (Armbruster et al. 2004). The inclusion of this methodologies have also provided of conceptual advances in our appreciation of phenotypic evolution (Garland and Adolph 1991; Felsenstein 2004): 1) phenotypic variation cannot be considered always adaptive; 2) the inclusion of phylogeny as a hierarchical dependency has improve the fit of the models of phenotypic evolution; 3) model of character evolution has become more general (parametrized) and complex; 4) phylogenetics and comparative biology have been integrated and provide a more accurate description of the phenotypic evolution; 5) reconstruction of ancestral phenotypes have been possible through the incorporation of phylogenies; and 6) experimental evolution has been enhanced by providing a temporal description based on hierarchical structure of phylogenetic inferences.

Some researchers have been resistant to include phylogenetic information in comparative analyses since the first statistical methodologies were established in the late 1980's (e.g., Felsenstein 1985, Cheverud et al. 1985b). This gave rise a long debate about the inclusion or not phylogenetic information if the characters in questions have low phylogenetic information (Westoby et al. 1995a, b, c; Björklund 1997). Several reasons usually account for the criticisms against the inclusion of phylogenetic analyses (Garland et al. 2004). First, all phylogenetic comparative analyses require phylogeny and this is especially difficult from some poorly known taxonomic groups (e.g., cnidarians and metazoan parasites) (Hillis 1998). Second, a reasonable number of taxa (e.g., > 20) should be included in the comparative analysis to get a robust inference statistical inference with narrower confidence intervals (Garland et al. 1999). Third, the models of character evolution are usually more complex that a random walk with constant variance (i.e., Brownian motion model) (Harvey and Pagel 1991; Losos 1999). Fourth, significant measurement error in the estimation of the population parameters will introduce biases in the comparative analyses, which is usually dependent on the experimental design and not in the phylogenetic reconstruction (Garland et al. 2004). Fifth, comparative analyses cannot per se infer causation as they are correlational in nature and the inferences are strictly based on the interpretation by the researcher (Harvey and Purvis 1991). Finally, some immediate ecological factors such as climate are more likely affect trait variation and they are not of evolutionary origin (Westoby et al. 1995a, b).

In spite of all these limitations, comparative analyses still provide a better fit to biological data. Some even regard to the standard non–phylogentic statistical analyses as special cases of the phylogenetic methodologies without hierarchical relationship among taxa (i.e., star phylogeny with equal branch lengths among terminals) (Blomberg et al. 2003). Simulation and empirical analyses have demonstrated that the inclusion of

phylogenetic information provides a similar or better fit to the data than non–phylogenetic methods (Rohlf 2001; Freckleton et al. 2002). Some authors have recommended that results of both phylogenetic and conventional analyses should be presented (Price 1997; Garland et al. 1999). Additionally, several test for degree of phylogenetic signal and diagnostics have been developed (e.g., Pagel's λ, and Blomberg's *K*) (Pagel 1999; Blomberg et al. 2003).

The use of phylogenetic comparative analysis has expanded beyond the simple correlational or ancestral trait reconstruction scheme. Several new methodologies based trait evolution and phylogenetic reconstructions have been developed. For example, diagnostics of taxonomic bias and its contribution to comparative analyses of common variables (e.g., metabolic rates) can be now incorporated as weighted parameters (Lajeunesse 2009). The strength of selection and drift can also be estimated from phylogenetic comparative analyses providing of model testing of neutral to adaptive hypothesis (Lajeunesse 2009). Moreover, we can determine the expected direction of phenotypic change based on estimated magnitude of natural selection or drift. Analyses of character association and the rates of speciation and extinction in a lineage can also be inferred from the association of phylogeny and character states (Maddison et al. 2007). Conservation priorities can also be derived from the association of ecological traits, phylogeographic information, and proyected climatological change (Redding and Mooers 2006; Isaac et al. 2007; Steel et al. 2007). Therefore, the extension of early comparative analyses into more complex models of phenotypic evolution under a phylogenetic framework will bring light to the process of biotic diversification.

Chapter 3: Methods to Incorporate the Effect of Shared Ancestry (Phylogenetic Signal)

The earliest methods to include phylogenetic information were based on taxonomic ranking, classification hierarchy, and pairwise comparisons (Felsenstein 1985; Harvey and Pagel 1991; Garland et al. 2004). These methods try to maximize species phylogenetic distinctiveness by defining groups based on evolutive affiliation (e.g., comparing two taxa of well-established and distinct families). However, these methods are only intuitive and the failure to incorporate an estimate of phylogenetic relationships introduces a pseudoreplication bias and increases type I error (Harvey and Pagel 1991).

Since the late 1980s, several comparative methods were developed to incorporate explicitly phylogenetic information under a statistical framework. A list of the most relevant include: phylogenetic independent contrasts (PIC) (Felsenstein 1985), phylogenetic generalized least squares (PGLS) models (Grafen 1989; Martins and Hansen 1997; Garland and Ives 2000), Monte Carlo trait evolution simulations (Martins et al. 1991; Garland et al. 1993), phylogenetic autocorrelation (Cheverud et al. 1985a), generalized estimating equations (Paradis and Claude 2002), phylogenetic mixed models (Housworth et al. 2004), concentrated changes test (Maddison 1990), and phylogenetically structured environmental variation methods (Desdevises et al. 2003). However, the two extensively validated methods for continuous characters are PIC and PGLS (Rohlf 2001; Felsenstein 2004; Garland et al. 2004) and I will detail them in this report.

PHYLOGENETIC INDEPENDENT CONTRASTS (PIC)

The inclusion of phylogeny in comparative analyses was probably started after the publication 'Phylogenies and the comparative method' (Felsenstein 1985). The rationale of this publication was that the mainstream research in biology at that time only used conventional statistical analyses. Felsenstein and others recognized that most of such analyzes violated two fundamental assumptions: independence of observations and residual errors (Harvey and Pagel 1991). PIC provided a method to incorporate information of phylogeny (i.e., hierarchical structure of species relationships based on cladogenesis), branch lengths (i.e., estimates of genetic or temporal distance for the branching pattern), and a model of character evolution (Figure 3). However, the method included some requirements if it were applied to its best: (1) a complete bifurcating pattern in the phylogeny, (2) continuous measurements of traits, and (3) a model of character evolution that follow an additive stochastic change (i.e. Brownian motion) (Blomberg et al. 2003).

The PIC algorithm goal is to estimate independent measurements (contrasts) that can be used to inferences using more conventional statistical analysis (Felsenstein 2004). PIC uses the information of actual species (i.e., tips of phylogeny), it goes down to each internal node in a completely bifurcating tree (Rohlf 2001). PIC procedure starts at the nodes before the tips and contrasts (measures the difference) in a random variable between its two daughter branches. The procedure is iterated until the contrast for the root node is found. The PIC contrast at each node depends on the scores of the contrasts of the nodes or tips one level above, which makes this algorithm recursive. The number of contrasts estimated after the procedure is n–1 when n is the number of tips (usually species, lineages, or individuals). Overall, PIC procedure will transform the raw data into of corresponding contrast that have null phylogenetic covariance with equal variance

(rate of character change), which makes the contrasts independent and standardized. The PIC contrasts can also be used for multivariate statistical methods, which include factor analysis (e.g., PCA and PAF) and multivariate regression (Harvey and Pagel 1991; Garland et al. 1992).

PIC algorithm can be illustrated using vector algebra (Rohlf 2001) or following the original description by Felsenstein (Felsenstein 1985; Felsenstein 2004). For this report, I will summarize Felsenstein's original description:

PIC Assumptions:

The phylogeny is fully bifurcating (i.e., polytomies should be resolved). The phylogeny has positive branch lengths (i.e., problematic in branch lengths near zero). The model of character evolution has to be Brownian motion (i.e., problematic if characters evolve under stabilizing selection).

PIC Procedure description:

1) Given a completely bifurcating phylogeny with branch lengths of *i* species and each species has a continuous estimable of the phenotypes X*i* and Y*i*. The numeric differences of X between pairs of sister species can be calculated and they are independent. These conditions should also apply to phenotype Y.

2) Given the branch lengths represent estimates of v_i temporal units since the split of the sister species. Assuming a Brownian motion model of evolution, we expect that small character displacements in X*i* have occurred in a random walk (i.e., equally likely positive or negative) during the time v_i. Therefore, we expect total displacement of X*i* at a node should come from the distribution $\sim N(0, s_X^2)$ with s_X^2 proportional to v.

3) Given that X and Y might undergo different rates of character evolution and vary in their degree of relationship (i.e., correlated to independent). We expect that after one unit of time the total phenotypic change in X and Y will have a mean of 0 and s_X^2 and s_Y^2, respectively. Extending to the v units of time since the split among daughter lineages, the phenotypes of X and Y should have a mean of zero and their variance will be $s_X^2 v$ and $s_Y^2 v$, respectively.

4) We use the estimations of phenotype X and Y from pairs of sister lineages or nodes i and j with a common ancestor node k (Figure 4). In the case of phenotype X, sister taxa will have an expectation of zero and a variance of $s^2_X (v_i + v_j)$, where s^2_X is a common to all contrast, v_i is the time unit from ancestor k to node i, and v_j is the time unit from ancestor k to node j.

5) We compute a contrast $X_i - X_j$ that should have an expectation of zero and a variance proportional to $v_i + v_j$.

6) We remove the two tree tips (i and j) but preserving the ancestor node k that becomes a new tip. The value for the tip X_k is calculated as a weighted average based on the branch lengths of tips i and j (i.e., v_i and v_j, respectively) and the values of the phenotypes X_i and X_j in the following form:

$$X_k = \frac{(1/v_i)X_i + (1/v_j)X_j}{(1/v_i) + (1/v_j)} \qquad (1)$$

Therefore if $v_i = v_j$ are the same then $X_k = (X_i + X_j)/2$, if $v_i \neq v_j$ then X_k is a weighted averaged based on branch lengths v_i and v_j.

7) We adjust the branch length of tip k (i.e., previously node k) from v_k to $v_k + v_i v_j$ $(v_i + v_j)$. This adjustment is necessary because the weighted average of X_k from equation

(1) is only an estimate of the phenotypic value of the ancestor and introduces an error relative to the lengths of v_i and v_j. This adjusted estimate is a contrast that will be included in the statistical analysis and it is independent of the phylogeny.

8) The algorithm is iterated from steps 4 to 7 to generate $n - 1$ contrasts from original n tips of the tree. Each contrast can be finally standardized by dividing by the square root of s^2_X to bring all contrasts to a common variance for phenotype X. A similar process can be performed for phenotype Y (i.e., steps 4 to 7) to generate $n - 1$ contrasts and standardized by by dividing by the square root of s^2_Y.

9) We will have generated $n - 1$ independent contrasts for phenotype X and Y. However, contrasts for X and Y are not necessarily independent and the covariance can be estimated at any given node (except the root):

$$\mathrm{Cov}[X_i - X_j, Y_i - Y_j] = 2v_i s_X s_Y r_{XY} \qquad (2)$$

where, v_i is the measure in temporal units (i.e., branch lengths) from last common ancestor k, s_X is the standard deviation for phenotype X, s_Y is the standard deviation for phenotype Y, and r_{XY} is correlation coefficient between X and Y. It is important to know that X and Y do not necessarily have the same rate of character change (i.e., $s_X = s_Y$).

10) The step 8 has standardized the contrast that can be regarded as drawn independently from a bivariate XY~N (0 , s_X=1 s_Y=1) with a r_{XY} that is unknown. Therefore, a test of independence of X and Y is reduced to test if $r_{XY} = 0$ or not. This last step will provide a unbiased test of correlated change between phenotypes X and Y useful for comparative analyzes.

PHYLOGENETIC GENERALIZED LEAST SQUARES (PGLS).

The introduction of PGLS extended the approach of PIC, as this one is considered a special case (Rohlf 2001). The PGLS was originally introduced by Grafen (Grafen 1989), but it was explicitly described in the late 1990s (Martins and Hansen 1997; Pagel 1997). During this decade (i.e., 2000s), the PGLS methodology has been improved by including confidence intervals (Garland and Ives 2000), alternative algorithms for variance-covariance matrix calculation (Butler and King 2004), and Bayesian estimation of the regression parameters (Pagel et al. 2004; Pagel and Meade 2007).

The requirements of PGLS are similar to PIC and include the topology with branch lengths, but it does not calculate contrasts. PGLS involves a modified GLS analysis with error terms that are not assumed to be independent nor identically distributed (Garland and Ives 2000). Therefore, branch lengths and tree topology are used to calculate a corrected variance-covariance matrix of the error terms that are independent of the phylogenetic inertia (Freckleton et al. 2002). PGLS assumes a Brownian motion model of character evolution with the following expectations: (1) GLS estimates of the regression parameters are also maximum likelihood estimates, and (2) elimination of non-independence bias as the error terms are drawn from a multivariate normal distribution (Garland and Ives 2000). Moreover, the estimated PGLS variance-covariance matrix can be used for factor analyses (e.g., PCA and confirmatory factor analysis), and ancestral state reconstruction (Lajeunesse 2009).

PGLS algorithm can be described in terms of regression with phylogenetically corrected data (Garland and Ives 2000). For this report, I will summarize the approach of Garland and Ives' (Garland and Ives 2000) modified from Martins and Hansen (Martins and Hansen 1997).

PGLS PROCEDURE DESCRIPTION:

1) Let y_i (dependent) and x_i, (independent) denote continuous measurements of two variables (phenotypes) from a species *i* from a dataset of *n* species in the phylogeny. We can describe the following regression equation:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad (3)$$

where β_0 and β_1 are regression coefficients and ε_i is an error term with mean zero but not independent from other similar terms. The values for ε_i are correlated among different species, which is derived from shared phylogenetic history. Therefore, phylogenetically closer species are expected to be more similar for y_i and x_i phenotypic values than more distantly related species.

2) The PGLS algorithm is implemented by dealing explicitly with the correlations among the ε_i for all extant taxa. So, a general equation from (3) can be written as:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad (4)$$

where **Y** is a vector of the values of the dependent y_i of the phenotype Y with *n*-dimensions (i.e., *n* number of extant taxa in the tree). **X** is a *n* x 2 matrix of values of the of the dependent x_i values of the phenotype X whose first column consist of ones and the second of the x_i values. The parameter $\boldsymbol{\beta}$ is equal to $[\beta_0 , \beta_1]'$ (i.e., a transpose of the regression parameters). Finally, $\boldsymbol{\varepsilon}$ approximates to an *n*-dimensional multivariate normal distribution with mean zero and a variance-covariance matrix $\sigma^2\mathbf{C}$, where σ^2 is a scalar measuring the expected rate of character change under a Brownian motion model and **C** is a phylogenetic matrix. The elements of the **C** matrix include the information of the

phylogenetic relationship among taxa based on the tree topology and branch lengths. The description of the calculation of the variance-covariance matrix and the phylogenetic **C** matrix will be expanded in the next section of this report.

3) PGLS algorithm progresses by iterating different values on β until the estimated values of **Y** converge to the empirical values y_i in the extant species. Several methodologies have been proposed to estimate the best values of these parameters including a maximum likelihood estimators (Freckleton et al. 2002) and Bayesian estimators (Pagel and Meade 2007).

Chapter 4: The Variance–covariance Matrix Estimation while Incorporating Phylogenetic Signal

In order to compare multivariate data from biological organism, data has to be converted to integrate phylogenetic information (Garland et al. 1999). Several methods have been proposed to convert raw data into phylogentically independent measurements as indicated in the section 3 of this report. However, the underlying theme among comparative methods including PIC and PGLS is that they are both special cases of the generalized theory of least squares (OLS) (Martins and Hansen 1997; Garland and Ives 2000; Freckleton et al. 2002; Adams 2008; Lajeunesse 2009). Many common parametric statistical tests are also based on OLS (e.g., multiple regression, ANOVA, ANCOVA, PCA), which provide an interesting link to phylogenetic comparative methods. However, OLS has assumptions that must be satisfied in order to be unbiased: (1) model has to be linear, (2) independent variables do not have to be collinear, (3) data points have to be randomly sampled from a population, (4) measurement errors have to be independent and normally distributed, and (5) the independent variables share a common sampling variance (i.e., they are homoscedastic). Most phylogenetic methods based on OLS have to overcome two critical violations to the assumptions of OLS in order to be unbiased, which are the lack of common variance among independent variables and that the error are non-independent due to share ancestry of the taxa that were measured.

The problems of the phylogenetic correction under OLS framework are mostly based on equation (4) and most methodologies try to incorporate phylogentic dependency on the error term ($\boldsymbol{\varepsilon}$). In a non-phylogenetic OLS problem, we expect that $\boldsymbol{\varepsilon}$ is a scalar variance-covariance matrix (i.e., a diagonal matrix whose diagonal elements all contain the same scalar variance σ^2, $\boldsymbol{\varepsilon} = \sigma^2\mathbf{I}$). However, the phylogenetic inertia causes $\boldsymbol{\varepsilon}$ to be different than $\sigma^2\mathbf{I}$ and the off-diagonal elements are different than zero. Therefore,

corrections for phylogeny are based on modeling the variance-covariance matrix (usually named $\boldsymbol{\Sigma}$ matrix) that weights dependency based on phylogenetic relatedness. The $\boldsymbol{\Sigma}$ matrix contains in its main diagonal the common variance σ^2 (or adjusted weights to compensate for unequal size effects) and in its off-diagonal elements (i.e., covariances) the accounts of phylogenetic relationship among taxa. Therefore, the correction will allow to give less weight to closely related taxa when fitting a regression line through their data (Pagel 1997, 1999; Lajeunesse 2009).

CALCULATING $\boldsymbol{\Sigma}$ MATRIX

The elements of $\boldsymbol{\Sigma}$ matrix have important implications in how to model the character evolution and the necessary corrections for phylogenetic signal. The elements of main diagonal are the common variance of the independent variables of the taxa included in the OLS analysis. Recent comparative analyses have proposed to include a penalization in the elements of the main diagonal to account for inaccurate estimates among taxa (i.e., unequal variances) and specifically for phylogenetic meta-analysis (Adams 2008; Lajeunesse 2009). However, for this report I will focus in the off-diagonal elements of the $\boldsymbol{\Sigma}$ matrix, which are relevant to phylogenetic factor analyses.

The elements of the off-diagonal of the $\boldsymbol{\Sigma}$ matrix are weighted covariances modeled after the phylogenetic information among taxa. These covariances measure the correlation phylogenetic history based on the $\mathbf{C}$ correlation matrix of equation (4). Most comparative methods based on OLS often consider a common variance (σ^2) among taxa, which reduces the $\boldsymbol{\Sigma} \propto \mathbf{C}$ (Rohlf 2001). Therefore, for phylogenetic factor analyses the variance-covariance $\boldsymbol{\Sigma}$ matrix can be effectively reduced to estimate the $\mathbf{C}$ matrix (Pagel 1999).

The **C** matrix contains the correlations among taxa based on the phylogeny used to correct for shared evolutionary history (Garland et al. 1999; Lajeunesse 2009). The usual account of shared phylogenetic history is based on the branch length (*b*) distance between taxa estimated in the phylogeny. Therefore, **C** matrix of *k* taxa has in its main-diagonal a function of the total branch lengths for each taxa (i.e, b_i^{total} total distance from the root to the *i* taxon tip) and in its off-diagonal a function of the shared distance between contrasted taxa (i.e., b_i^{shared} distance from root between *i* and *j* taxa).

Several models of character evolution can be incorporated as modifications of the **C** matrix. I will use the notation of Lajeunesse (Lajeunesse 2009) and focus on two of the most common models: Brownian motion ($\mathbf{C}^{\mathbf{BM}}$ matrix) and the Ornstein-Uhlenbeck ($\mathbf{C}^{\mathbf{OU}}$ matrix).

The basic model is the $\mathbf{C}^{\mathbf{BM}}$ matrix, which assumes random character proportional to the branch lengths (i.e., time since divergence). Therefore, $\mathbf{C}^{\mathbf{BM}}$ matrix will include only the following elements:

$$C_{i,j}^{BM} = \begin{cases} b_i^{\text{total}} & \text{if } i = j \\ b_{i,j}^{\text{shared}} & \text{if } i \neq j \end{cases} \qquad (5)$$

where, b_i^{total} is total branch length distance from the root to the tip of *i* taxon and $b_{i,j}^{\text{shared}}$ is shared branch length distance from the root to the node before the split between *i* and *j* taxa (Rohlf 2001; Lajeunesse 2009).

The Ornstein-Uhlenbeck model (Hansen and Martins 1996) is the $\mathbf{C}^{\mathbf{OU}}$ matrix, which assumes an exponential relationship between contrasted taxa based on the phylogentic distance and a selection parameter (β). Therefore, $\mathbf{C}^{\mathbf{OU}}$ matrix will include only the following elements (Hansen 1997):

$$C_{i,j}^{OU} = \begin{cases} (1 - e^{-2\beta b_i^{\text{total}}}) / 2\beta & \text{if } i = j \\ (e^{-2\beta (b_i^{\text{total}} - b_{i,j}^{\text{shared}})} - e^{-2\beta b_i^{\text{total}}}) / 2\beta & \text{if } i \neq j \end{cases} \quad (6)$$

where, b_i^{total} is total branch length distance from the root to the tip of i taxon, $b_{i,j}^{\text{shared}}$ is shared branch length distance from the root to the node before the split between i and j taxa, and the selection parameter β that vary from neutral ($\beta \rightarrow 0$) to strong ($\beta \rightarrow \infty$) selection (Hansen 1997; Lajeunesse 2009). Interestingly, the $\mathbf{C}^{OU}$ matrix can approximate to $\mathbf{C}^{BM}$ matrix if the selection parameter $\beta \rightarrow 0$, which makes the Brownian motion a special case of Ornstein-Uhlenbeck model (Butler and King 2004).

Finally, all methods of phylogenetic comparative analysis assume a direct relationship with time expressed as branch lengths. Ultrametricity (i.e., all tree tips are contemporaneous in time) is required to be met in order to apply the model of character evolution and in the calculation of the **C** matrix (Garland et al. 2004). However, the timescale of the chronograms are not required to be absolute and corrections can be applied to phylogenies to make them ultrametric(Lajeunesse 2009). Therefore, a standardization of the **C** matrix to meet ultrametricity is necessary and can be performed by dividing elements b_i^{total} (Pagel 1994), or alternative corrections with other elements of control (Lajeunesse 2009).

USING C MATRIX IN STANDARD LINEAR REGRESSION

The **C** matrix can be used to solve the problem of correlated errors due to phylogenetic signal by making the error terms uncorrelated. Garland and Ives (Garland and Ives 2000) proposed an interesting application of matrix algebra to the problem of

phylogenetically correlated error terms. The method can be summarized as follows (Garland and Ives 2000): (a) **C** matrix is a real symmetric nonsingular matrix as described above; (b) there should exists another matrix **D** such that **DCD'** = **I** (i.e., an identity *n* x *n* matrix); (c) the **D** matrix can then be used to transform values of the traits (e.g., *x* and *y*) and the error term ε, such that **Z** = **DY**, **U** = **DX**, and $\boldsymbol{\alpha}$ = **D**$\boldsymbol{\varepsilon}$, which reduces equation (4) to

$$\mathbf{Z} = \mathbf{U}\boldsymbol{\beta} + \boldsymbol{\alpha} \qquad (7)$$

where, the new variance-covariance matrix of $\boldsymbol{\alpha}$ equal to $\sigma^2\mathbf{I}$ (i.e., no covariance elements are present), (d) this is true by the following: $V\{\alpha\} = E\{\alpha\alpha'\} = E\{D\varepsilon(D\varepsilon)'\} = E\{D\varepsilon\varepsilon' D'\} = \mathbf{D}E\{\varepsilon\varepsilon'\}\mathbf{D'} = (\mathbf{D}\sigma^2\mathbf{C})\mathbf{D'} = \sigma^2\mathbf{I}$, and (e) $\boldsymbol{\alpha}$ matrix also approximates to normal distribution, as $\boldsymbol{\alpha}$ is a linear transformation of $\boldsymbol{\varepsilon}$. The matrices **Z** and **U** can be back-transformed to matrices **Y** and **X**. Moreover, Garland and Ives (Garland and Ives 2000) also provided extensions to more than two variables and important approximations GLS equation (4). Including, the regression coefficient $\boldsymbol{\beta}$ vector, $\hat{\beta}$:

$$\hat{\beta} = (\mathbf{U'U})^{-1}(\mathbf{U'Z}) = (\mathbf{X'C}^{-1}\mathbf{X})^{-1}(\mathbf{X'C}^{-1}\mathbf{X}) \qquad (8)$$

unbiased estimate of variance σ^2 with *N* variables, $\hat{\sigma}^2$:

$$\hat{\sigma}^2 = (\mathbf{Z} - \mathbf{U}\,\hat{\beta})'\,(\mathbf{Z} - \mathbf{U}\,\hat{\beta})/(n - N) = (\mathbf{Y} - \mathbf{X}\,\hat{\beta})'\mathbf{C}^{-1}\,(\mathbf{Y} - \mathbf{X}\,\hat{\beta})/(n - N) \qquad (9)$$

variance-covariance matrix of $\boldsymbol{\beta}$, $V\{\beta\}$:

$$V\{\beta\} = \sigma^2 (\mathbf{U}'\mathbf{U})^{-1} = \sigma^2 (\mathbf{X}'\mathbf{C}^{-1}\mathbf{X})^{-1} \qquad (10)$$

estimated variance-covariance matrix of $\boldsymbol{\beta}$, $s^2\{\beta\}$:

$$s^2\{\beta\} = \hat{\sigma}^2 (\mathbf{U}'\mathbf{U})^{-1} = \hat{\sigma}^2 (\mathbf{X}'\mathbf{C}^{-1}\mathbf{X})^{-1} \qquad (11)$$

estimates of mean responses of $\mathbf{Y}$, $\sum_{i=1}^{n} a_{ik}^2 \ \hat{Y}$:

$$\hat{Y} = \mathbf{D}^{-1}(\mathbf{D}\mathbf{X}\ \hat{\beta}) \qquad (12)$$

estimated variance-covariance matrix of mean responses of $\mathbf{Y}$, $s^2\{\hat{Y}\}$:

$$s^2\{\hat{Y}\} = \mathbf{D}^{-1}\, s^2\{\hat{Z}\} (\mathbf{D}^{-1})' = \mathbf{X}\, s^2\{\beta\}\, \mathbf{X}' \qquad (13)$$

Chapter 5: The Variance–covariance Matrix in Phylogenetic Exploratory Factor Analysis

Factor analysis (FA) is a data reduction technique used to identify a reduced number of latent variables underlying a set of covaring directly measurable variables (Kline 1994; Brown 2006). The uses of FA include the reduction in the redundancy in a set of intercorrelated variables, determination of the dimensionality (i.e., degree of multicollinearity), and model testing of alternative hypothesis of the observed variable covariance. In order to assess with high reliability a FA, multiple observable variables are usually used. Each observed variable includes a proportion attributed to latent variable of interest (reliability) and the rest from measurement error (e.g., other influencing latent variables). Therefore, the implementation of FA is designed to extract from each observed variable the proportion that is measuring the construct without the error.

The basic steps of FA can be summarized in data gathering, method of factor extraction, and interpretation (Brown 2006). First, data collection is fundamental for any inference and is based on the type of factor (i.e., latent variable) to be characterized, the population of interest, and completeness of the dataset obtained (Tabachnick and Fidell 2007). Second, the FA methods are the estimation of the proportion of variability of the observed variable between the factor and measurement error (Jolliffe 2002). Two common FA methodologies include: principal component analysis (PCA) and common factor analysis (PAF). Third, interpretation of the results based on previous knowledge (i.e., theory) and the determination of significant relationships (i.e., factor loadings) with the latent factors (Kline 1994; Brown 2006). For this report, I will focus on the PCA and how the phylogenetically independent variance–covariance matrix can be used for this type of factor analysis.

All FA methods use a common path analysis notation and include four elements (Kline 1994). First, the latent variables or factors that are inferred elements and account for the pattern of association among observed variables. Second, the indicator or observable variables are the elements that derive from empirical data and they are used to account for the underlying factors. Third, the factor loadings are the path connecting the indicator onto the factor and it is characterized by its magnitude and direction. Finally, the error term or unique variance of the indicator that is not explained by the factor. Thus, the performance of an indicator is derived by the combination of the factor loading and measurement error.

Factor analysis has two different approaches in data reduction. Exploratory factor analysis (EFA) determines all factor loadings (i.e., magnitude and direction of relationship) between each observed variable on all latent variables. PCA is an example of EFA and it is the most common FA method use in Biology (Quinn and Keough 2002; Gotelli and Ellison 2004). Confirmatory factor analysis (CFA) resembles EFA, but some factor loadings are set to zero (i.e., independent or not related) and it is in more model testing under theoretical grounds by the researcher. Therefore, CFA can be use to compare alternative hypothesis of factor-indicator relationship and EFA is more exploratory and does not assume any a priori mode of relationship among factor and indicator variables. I will cover CFA under the following chapter of this report.

Principal component analysis (PCA) is method based on the assumption that given a set of n observable variables, in n-dimensional space, will have unevenly distributed variation and concentrated within few dimensions (Kline 1994; Brown 2006). Therefore, PCA will redistribute correlations among variables into perpendicular (i.e., orthogonal or uncorrelated) dimensions. The variance explained by the new dimensions or components is distributed in a decreasing order. Each component is a linear

combination of all *n* variables and the first component will explain the most of the simultaneous variation among all observed variables. From the remaining variation, the second component will explain the most and so on until we extract *n* components of *n* variables have explained all variation in the set. Therefore, PCA summarized the observed variability among a set of *n* variables in a set of *n* orthogonal components with most of the variation explained by the first components extracted.

One important feature of PCA is that it can be calculated using matrix algebra from a regression matrix **R** (Jolliffe 2002). For instance, for a given set of *n* variables ($\boldsymbol{X_1}$, **… ,** $\boldsymbol{X_n}$) can be replaced by *n* orthogonal components ($\boldsymbol{C_1}$, **… ,** $\boldsymbol{C_n}$) , given the following conditions about their linear combination for a given component $\vec{c}_k$:

$$\vec{c}_k = a_{1k}\,\vec{x}_1 + a_{2k}\,\vec{x}_2 + \ldots + a_{nk}\,\vec{x}_n \qquad (14)$$

where, every a_{ik} is a linear coefficient and $\vec{x}_i$ is the variable vector. Such that each $\vec{c}_k$ is an eigenvector described as coordinates of the original variable space ($\vec{x}_1, \ldots, \vec{x}_n$) under the condition:

$$a_{1k}^2 + a_{2k}^2 + \ldots + a_{nk}^2 = \sum_{i=1}^{n} a_{ik}^2 = 1 \qquad (15)$$

Therefore, in order to calculate a full PCA (i.e., all components or eigenvectors) for the correlation matrix **R** of *n* variables, we need to fulfill the following condition:

$$\mathbf{R}(\vec{c}_1, \ldots, \vec{c}_n) = (\lambda_1\vec{c}_1, \ldots, \lambda_n\vec{c}_n) \qquad (16)$$

where **R** is the correlation matrix, $\vec{c}_i$ is the components or eigenvectors, and λ_i are the eigenvalues. This relationship is important for interpretation as eigenvectors for a component are the coordinates of the component in the original variable space and eigenvalues are the variance explained by the component. Therefore, the equation (16) can be rewritten using a matrix notation where **P** is a matrix of eigenvectors and **Λ** is a diagonal matrix of eigenvalues:

$$\mathbf{RP} = \mathbf{P\Lambda} \qquad (17)$$

this can be rearranged in order to calculate the principal component equation:

$$\mathbf{P^{-1}RP} = \mathbf{\Lambda} \qquad (18)$$

Factor loadings are the correlation coefficients between the observed variables and components. Important information can be obtained from factor loadings including: (1) the proportion of variance in the variable that is explained by a component calculated as the square of each variable factor loading, and (2) the communality or the total variance in observed variable accounted for by all the factors, which is the sum of the squared factor loadings for all factors for a given variable. Therefore, we can manipulate equation (18) to obtain factor loadings such as

$$\mathbf{P^{-1}RP} = \mathbf{\Lambda} \rightarrow \mathbf{R = P\Lambda P'} \qquad (19)$$

using the property of orthogonal matrices $\mathbf{P' = P^{-1}}$. In order to calculate factor loadings, we can modify equation (19) to obtain factor loadings using the following property:

$$\mathbf{P\Lambda P' = (P\ \sqrt{\Lambda}\ )\ (\sqrt{\Lambda}\ P') = R} \qquad (20)$$

and factor loadings are $\mathbf{B = P\ \sqrt{\Lambda}}$ then:

$$\mathbf{P\Lambda P' = BB'} \qquad (21)$$

Alternatively, PCA con be understood from a multiple regression context. We can define a set of variables set of *n* variables ($\boldsymbol{X_1, \ldots, X_n}$) such that the weighted sum of all variables give *n* orthogonal components ($\boldsymbol{C_1, \ldots, C_n}$) in the form:

$$c_k = a_1 X_1 + \ldots + a_n X_n \qquad (22)$$

with the condition that

$$a_1^2 + \ldots + a_n^2 = \sum_{i=1}^{n} a_i^2 = 1 \qquad (23)$$

under PCA will try to maximize the sum of the squared coefficients on the sum c_k and each variable X_i.

Many other calculation and inferences can be done after the PCA has been completed including component rotation or the number of components to retain. However, they are beyond the scope of this report and I will suggest to potential readers to review appropriate literature (e.g., Jolliffe 2002).

USING THE PHYLOGENETICALLY CORRECTED VARIANCE–COVARIANCE MATRIX IN EXPLORATORY FACTOR ANALYSIS

Factor analysis as many other multivariate methods based on GLS methods assume that the multivariate error variances are independent (Garland et al. 1993; Jolliffe 2002). In case of biological data, this assumption is violated unless it is accounted in the analysis or excluded from the input correlation or variance–covariance matrices (Garland et al. 1993). The correction and calculation of the variance–covariance matrix after correction for phylogenetic signal was addressed in chapter 4 of this report (Figure 4). Therefore, I will indicate describe the calculation of the principal components using the phylogenetically corrected variance–covariance matrix.

The methodology of calculating PCA given a variance–covariance matrix can be summarized in the following steps (Jackson 2003): (1) calculate the correlation matrix from the variance–covariance matrix of *n* variables; (2) calculate the eigenvectors and eigenvalues from the correlation matrix; and (3) extract *n* components from *n* variables. Therefore, the only step that is significantly different from previous description of PCA is the conversion from a variance–covariance into a correlation matrix. In this report, I will describe how such conversion can be performed.

Given a phylogenetically corrected variance–covariance matrix **A** it can be transformed to correlation matrix **R** (Jolliffe 2002). It can be transformed as follows: (1) calculate the $\mathbf{A}^{-1}$, (2) get the diagonal elements of $\mathbf{A}^{-1}$ into diagonal matrix **D,** (3) calculate the square root of the elements of matrix **D**, (4) compute correlation matrix **R** by calculating **DAD**. The correlation matrix can then be used in the PCA calculation described above.

Chapter 6: The Phylogenetic Confirmatory Factor Analysis and Structural Equation Modeling

Exploratory (EFA) and confirmatory (CFA) factor analyses share methodological similarities, but profound conceptual differences (Kline 2005; Brown 2006; Grace 2006). For instance, with EFA, we try to determine the number of factors and calculate the loadings on every factor by all available indicator variables. With EFA, we can specify the number of factors to be extracted and constrain the factors to be orthogonal or not (i.e., independent or correlated), but we cannot constrain the indicators variable loadings to be zero. We also do not have a previous idea of the structure of factors and indicators loadings; and the answer is unique for a given correlation matrix. With CFA, we can specify a priori several models (i.e., a factor structures) of the number of factors, loadings from indicator variables, and error terms in order to find the most parsimonious model (i.e., a simple structure) that best explains the empirical variance-covariance matrix.

CFA methodology can be summarized in the following steps and I will detail each part in the following paragraphs (Figure 4). First, we specify a model of factor and indicator variables loadings or correlations (usually in the form of a path diagram). Second, we determine if the model is identified (i.e., number of unknown parameters is equal or less than the known parameters) in order to estimate a unique set of parameters. Third, we estimate the model (i.e., run an algorithm to estimate the model parameters). Four, we determine the fit of the model to the empirical data (i.e., variance-covariance matrix) using fit indices. Fifth, we re-specify (i.e., modify) the model structure if it does not fit using theoretical evidence or modification indices and iterate starting from the second step until we get the best answer. Finally, we do the inference on the base of theoretical ground and additional empirical evidence (e.g., cross-validation).

Model specification is strictly derived from theoretical grounds or empirical evidence that include latent, indicator, and error term variables (Brown 2006). Latent factors are unobserved variables that are measured by set observable indicator variables. Observed variables or indicators are measurable from a given population and they present an aspect of the latent variable. Error terms are the part of the indicator variable not explained by the factor and it might come from other unseen factors or measurement error. With CFA, the researcher can propose a model based on the relationships between the set of latent factors and their indicator variables. These relationships are derived from empirical evidence or the researcher proposing a hypothesis on how factors and indicators are related to each other. In most CFA, the model propositions usually underline causative and correlation basis. For instance, the researcher can propose that factor-1 affects factor-2 on the basis of a temporal progression (e.g., age mother affects the viability of their gametes, and not vice versa). In other hand, the researcher can propose how indicators measure which factors, so he/she can constrain some loading to be zero (i.e., a proposition of negligible direct relationship between a factor and indicator). It is important that this is not the case for EFA, where all indicator loadings on each factor have to be estimated (Brown 2006).

A model is identified if the number of parameters to be estimated is less or equal to the number of known parameters in the model (Kline 2005; Brown 2006; Grace 2006). Under CFA, the known parameters are the number of elements (p^*) in the empirical variance-covariance matrix. The parameters usually estimated (q) are factor variances, error terms variances, covariances among factors or error terms, and factor loadings. Therefore, a simple equation to determine if a model is identified is the following:

$$p^* = \frac{p(p+1)}{2} \quad \frac{p(p+1)}{2} \qquad (24)$$

$$\text{model uniqueness} = p^* - q \qquad (25)$$

where p^* is the elements in the variance-covariance matrix, p is the number of indicator variables, and q is the number of estimated parameters. Therefore, a model is under-identified if we have more unknown than known parameters (i.e., $p^* < q$), and single model cannot be estimated for the empirical variance-covariance matrix. A model is just-identified if we have equal unknown and known parameters (i.e., $p^* = q$), and the model implied will recover exactly the empirical variance-covariance matrix. A model is over-identified if we have less unknown than known parameters (i.e., $p^* > q$), and the model implied would recover a variance-covariance matrix that is less fit than the empirical. However, the over-identified models are of interest, as they allow the inference of more parsimonious answers of the factor, indicator, and error term relationships.

In order to contrast different structural models, we need that they are over-identified. For instance, the researcher has to propose model of variable relationships by fixing some paths to specific values (usually one), dropping paths (e.g., covariance between factors equal to zero), and estimating the remaining parameters. Fixing an indicator loading or setting a variance of a factor to one usually achieves scaling a factor (Kline 2005; Brown 2006). However, most over-identified models are proposed on the base of theoretical grounds and specific hypothesis about the relationships among indicators and factors.

The estimation of a CFA has some similarities with EFA algorithms (Brown 2006). For instance, consider the indicator (observed) variables (X_1, … ,$X_{n)}$ can be derived from the following linear equation each variable:

$$X_i = \lambda_i \xi + \delta_i \qquad (26)$$

where λ_i is the factor loading from each X_i, ξ is the factor, and δ_i is the measurement error from each X_i. Then, consider the observed variance-covariance sample matrix $\boldsymbol{S}$ and the variance-covariance population matrix $\sum$. The relationship between the linear equation (26) and the elements of the matrices $\sum$ are equivalent by the following:

$$\sum = \boldsymbol{\lambda \phi \lambda' + \theta \delta} \qquad (27)$$

where $\boldsymbol{\lambda}$ is a vector matrix of factor loadings, $\boldsymbol{\phi}$ is a variance-covariance matrix of the ξ_i factors and $\boldsymbol{\theta \delta}$ is a diagonal matrix of the variances of the measurement errors. Therefore, the variance-covariance matrix among X_is or $\sum$ can be estimated from set of matrices $\boldsymbol{\lambda}$, $\boldsymbol{\phi}$, and $\boldsymbol{\theta \delta}$.

CFA as many other multivariate methods assume that the multivariate error variances are independent (Garland et al. 1993; Jolliffe 2002). In case of biological data, this assumption is violated unless it is accounted in the analysis or excluded from the input variance–covariance matrix $\boldsymbol{S}$ (Garland et al. 1993). The calculation of the corrected variance–covariance matrix after accounting for phylogenetic signal was addressed in chapter 4 and 5 of this report. Therefore, I suggest the reader to review both previous chapters and assume that the input variance–covariance matrix $\boldsymbol{S}$ has already been corrected for phylogenetic signal.

CFA and structural equation modeling (SEM) are a form of covariance analysis among the observed X_i variables (Grace 2006). Therefore, CFA and SEM are based on a minimization of a discrepancy function or loss function $\boldsymbol{E}$. This is achieved by minimizing the discrepancy between the variance-covariance matrix of the observed variables in the sample ($\boldsymbol{S}$) and the variance-covariance matrix implied for the population ($\sum$) by the model proposed by the researcher. The discrepancy function is defined as

$$\boldsymbol{E} = \boldsymbol{S} - \sum(\boldsymbol{\theta}) \qquad (28)$$

where $\boldsymbol{S}$ is the observed variance-covariance matrix, and $\sum$ is the model implied variance-covariance matrix using a set of $\boldsymbol{\theta}$ parameters for path loadings, covariances, and error terms. However, the elements of $\sum(\boldsymbol{\theta})$ are unknown and they have to be estimated by the $\boldsymbol{S}$ matrix. Therefore, the estimates of the model parameters ($\hat{\theta}$) are included in the $\sum(\boldsymbol{\theta})$ matrix resulting in a implied variance-covariance matrix, $\sum(\hat{\theta})$. Therefore, the discrepancy function

$$\boldsymbol{E} = \boldsymbol{S} - \sum(\hat{\theta}) \qquad (29)$$

is used to estimate the unknown parameters of the structural model. Analytically, $\boldsymbol{E}$ is minimized by setting starting values of the $\hat{\theta}$ parameters and iterated until the $\boldsymbol{E}$ converges. The value of $\boldsymbol{E}$ is estimated with each set of new parameters and the fit of the implied model is determined. The criterion for convergence is predetermined to a minimal discrepancy (difference) threshold value (e.g., < 0.001) of $\boldsymbol{E}$ before its acceptance. Usually, multiple independent runs with different starting values are proposed and the expectation is that all of them converge to similar $\hat{\theta}$ parameters.

Therefore, CFA and SEM are method to test how well proposed models account for the variance-covariance matrix of the observed variables.

Several methods have been proposed to minimize the discrepancy function $\boldsymbol{E}$ (Grace 2006). Some of the most important are: Maximum Likelihood (ML), Unweighted Least Squares (ULS), and Generalized Least Squares (GLS). ML approach minimizes the sum of the squared differences for $\boldsymbol{S} - \sum(\hat{\theta})$ after weighting them by the inverse of the $\sum(\hat{\theta})$. ULS approach minimizes only the sum of the squared differences for $\boldsymbol{S} - \sum(\hat{\theta})$. GLS approach minimizes the sum of the squared differences for $\boldsymbol{S} - \sum(\hat{\theta})$ after weighting them by the inverse of the $\boldsymbol{S}$. All methods provide a χ^2 statistic based on the difference of $[(\boldsymbol{S} - \sum(\hat{\theta})]^2$ in the form of

$$\chi^2 = E_{\mathrm{ML}}\,(n-1) \qquad (30)$$

where E_{ML} is the minimized fitting function and n is the sample size of the sample population. The significance of the χ^2 statistic is determined by the degrees of freedom or the number of parameters estimated (i.e., $p^* - q$). The structural equation model is rejected if χ^2 more than a critical value (c_α) and α significance level.

Structural model fit is reported by providing overall model fit (i.e., χ^2 goodness of fit) and several fit indices (Brown 2006). Large sample sizes tend to cause significant χ^2 statistics as implied from equation (30). Therefore, other fit indices provide alternative evaluation of model fit and they are classified in incremental and absolute fit indices. Incremental fit indices measure the proportional improvement of a proposed model against a null model of all indicators as uncorrelated variables. The most common incremental indices are the Normed Fit Index (NFI) (Bentler and Bonnett 1980), Tucker-Lewis Index (TLI) (Tucker and Lewis 1973), and the Comparative Fit Index (CFI)

(Bentler 1990). Absolute fit indices measure the accuracy of the structural model in reproducing the observed variance-covariance matrix. The most common absolute indices are the Standardized Root Mean Square Residual (SRMR) (Joreskog 1993) and the Root Mean Square Error of Approximation (RMSEA) (Steiger and Lind 1980). The determination of good fit based on the indices reported are > 0.95 for NFI, CFI, and TLI; < 0.10 for SRMR; and < 0.06 for RMSEA (Hu and Bentler 1999). For this report, I have not included the derivations of the fit indices, but it is encouraged that readers refer to literature cited.

Structural model does not have an adequate fit, if the model has a significant χ^2 statistic and evidence of poor fit based on the indices described above. However, the model can be re-specified (i.e., modified) and subsequently re-tested for fit. The basis of model modification has to have significant theoretical bases by proposing or dropping parameter estimation. Two common model re-specification techniques are the Lagrange Multiplier and Wald statistic modification index algorithms (Grace 2006). Lagrange Multiplier (LM) is an iterative approach where the path and parameters are added to the model and fit improvement is determined by statistical significant changes in the overall χ^2 statistic. Parameters are progressively added until the fit model of the model is not significantly better than without the added parameter. The Wald statistic is also an iterative approach where the path and parameters are dropped from the model and fit improvement is determined by statistical significant changes in the overall χ^2 statistic. Parameters are progressively dropped until the fit model of the model is significantly worst than with the dropped parameter. The LM and Wald statistic approaches usually have several independent starts and an overall consensus for adding or dropping specific parameters is presented to the researcher.

The final step is the interpretation of the accepted model in the context of theory and the population where the data came from. Several aspects have to be considered after a model is favored. First, the true model almost never be recovered with enough accuracy, but the structural model is the best approximation of the truth. Second, the preferred model is one of multiple equivalent models that equally explain our data and we shall be willing to propose alternatives. Third, the model preferred was derived from a given dataset and cross-validation with independent datasets is always desirable. Finally, the model is a simplification of the world and we have to assume that the complexity is always a reality.

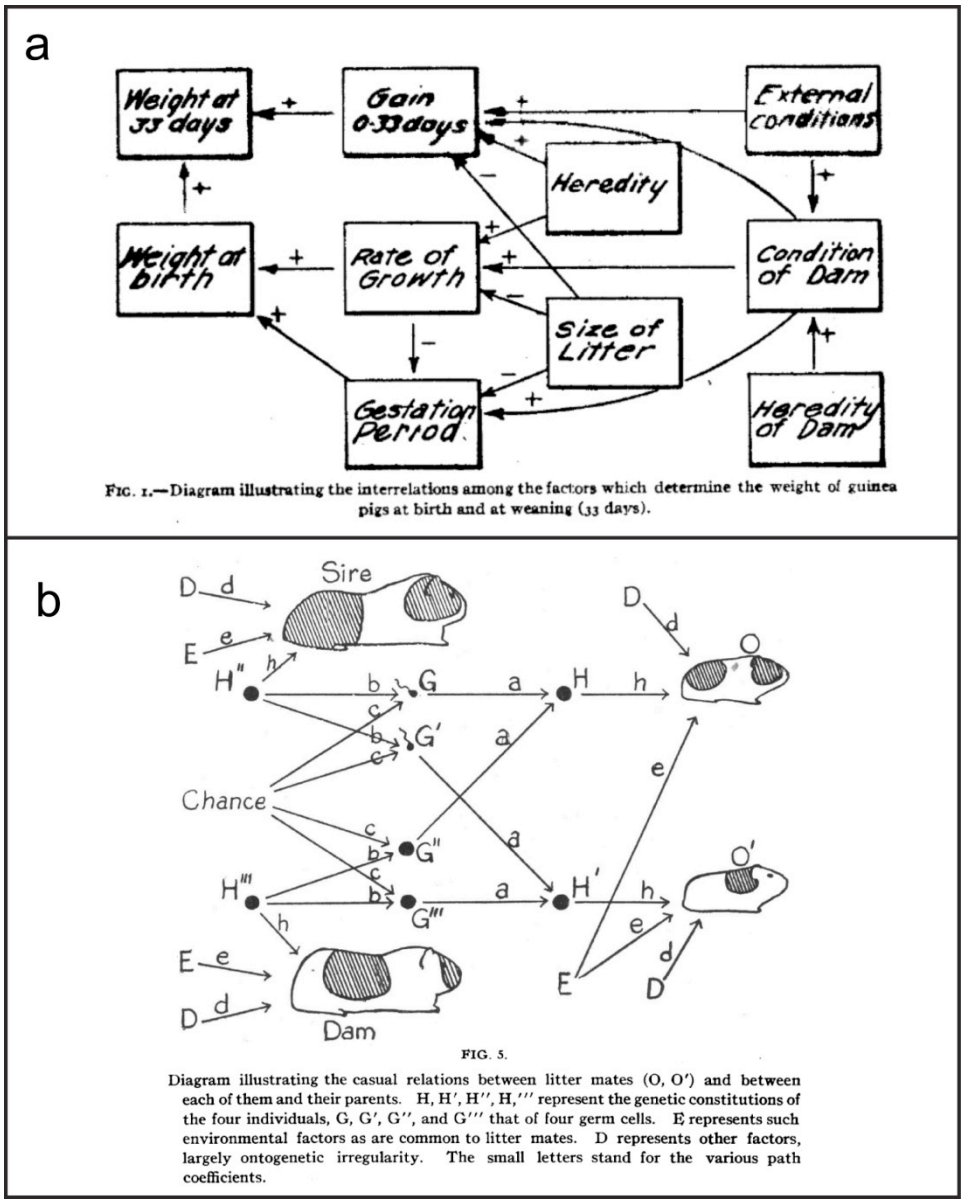

FIG. 1.—Diagram illustrating the interrelations among the factors which determine the weight of guinea pigs at birth and at weaning (33 days).

FIG. 5.

Diagram illustrating the casual relations between litter mates (O, O′) and between each of them and their parents. H, H′, H″, H,‴ represent the genetic constitutions of the four individuals, G, G′, G″, and G‴ that of four germ cells. E represents such environmental factors as are common to litter mates. D represents other factors, largely ontogenetic irregularity. The small letters stand for the various path coefficients.

Figure 1: Sewell Wright path diagrams depicting the fitness, reproduction, and growth traits interrelationships of guinea pigs. (a) Wright's 1921 paper (Wright 1921) on causation and correlation shows the negative or positive relationships between hereditary factors and measurable variables (e.g., rate of growth and gestation period). (b) Wright's 1920 paper (Wright 1920) on the importance of hereditary and environmental variables on phenotypic traits of guinea pigs. Wright introduced the concept of path analysis and the some of its rules to address correlation by diagrams as early as the 1920's.

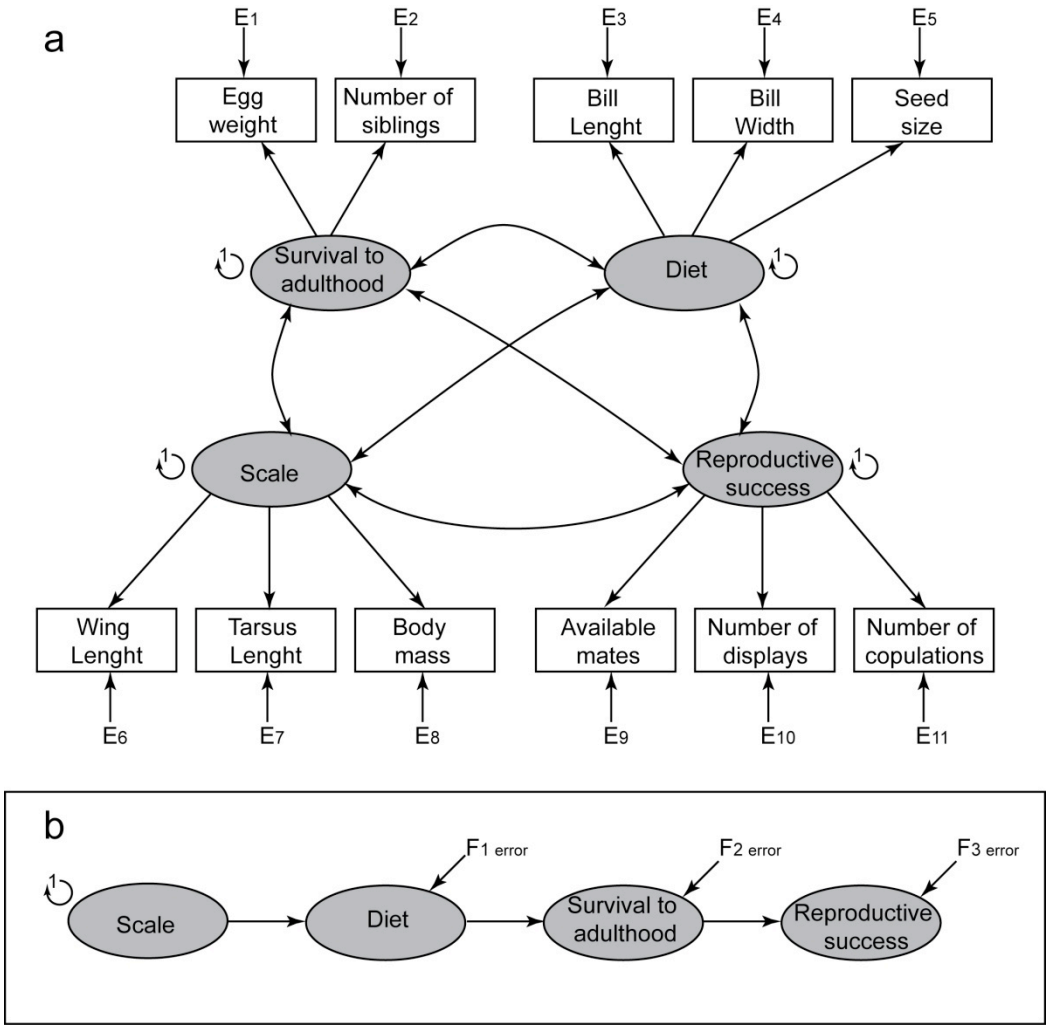

Figure 2: An exemplar measurement (a) and structural model (b) of observed and latent variables related to the reproductive success based on data of Darwin's finches (Grant 1999). (a) The measurement model depicts the latent (implied, grey ovals), their indicator (observed, white boxes), and error variables. Single-headed arrows indicate direct effect from latent or error on indicator variables. Double-headed arrows indicate correlation. The circles with 1 indicate that the model has been standardized. (b) One of the proposed structural models of relationships among latent variables. The model indicated suggest a direct effect of scale on diet, diet on survival to adulthood, and survival on reproductive success. The model (b) is illustrative and it does not reflect an actual SEM analysis performed in Darwin's finches data.

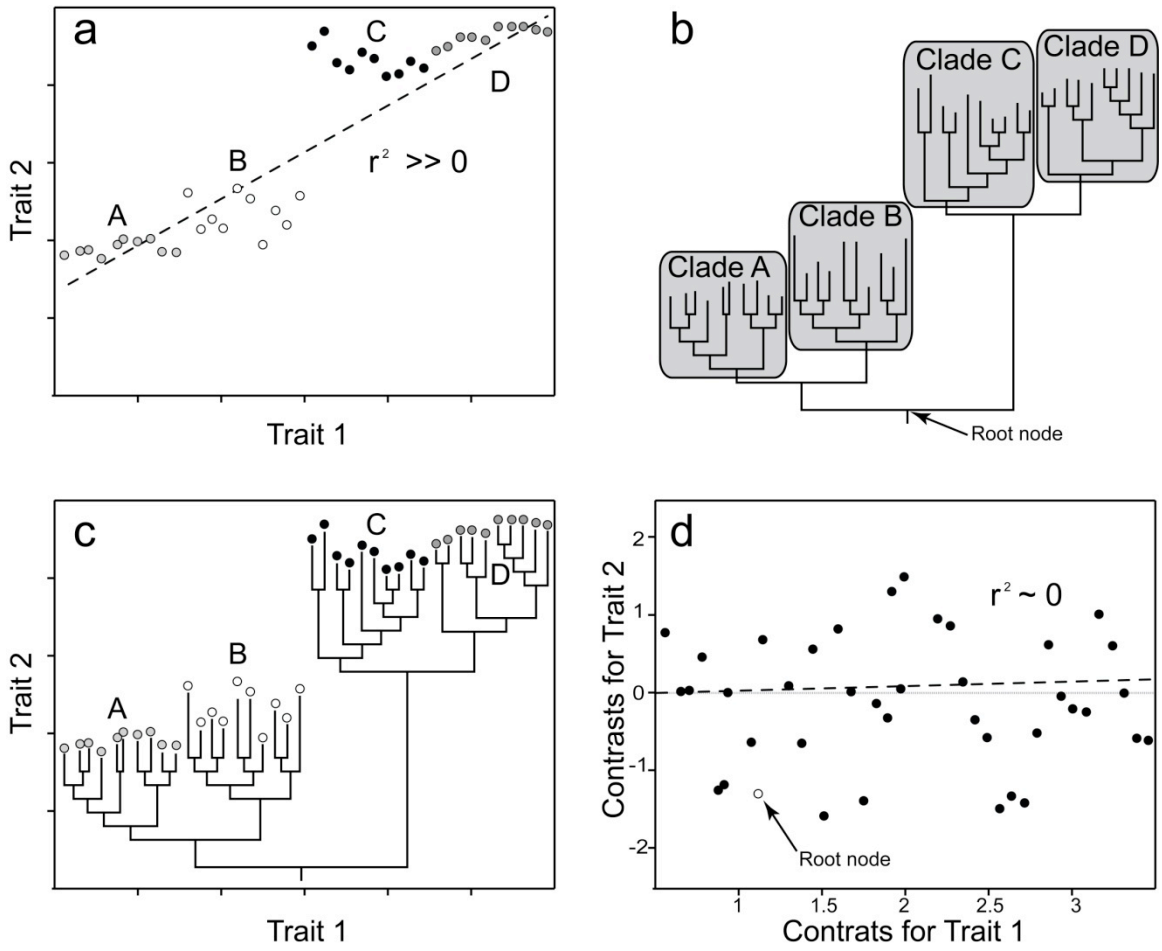

Figure 3: An exemplar use of independent contrasts with two traits that have evolved uncorrelated in 40 species within four lineages (Clades A-D). (a) Plot of the bivariate values of trait 1 versus 2 of all species. The dashed line is the linear regression with a significant positive correlation value between both traits. (b) Phylogeny of the 40 species with their lineage affiliation. Species in Clade A are evolutionarily more closely related to Clade B, than species in Clades C and D. (c) Plot of the bivariate relationships with the phylogeny superimposed, the reader might notice that data points are not independent and there is a hierarchical nature of the relationships among species. We might need to account or correct for the phylogenetic signal before statistical analyses can be performed if independence of data points is an assumption (i.e., multivariate regressions, MANOVA, factor analysis, and structural equation modeling). (d) Plot of the phylogenetic independent contrasts (PIC) for the two traits, the number of contrasts is 39 (i.e., $n - 1$ species). The regression line is forced through the origin (Felsenstein 1985) and we reject the association between both traits after phylogenetic correction.

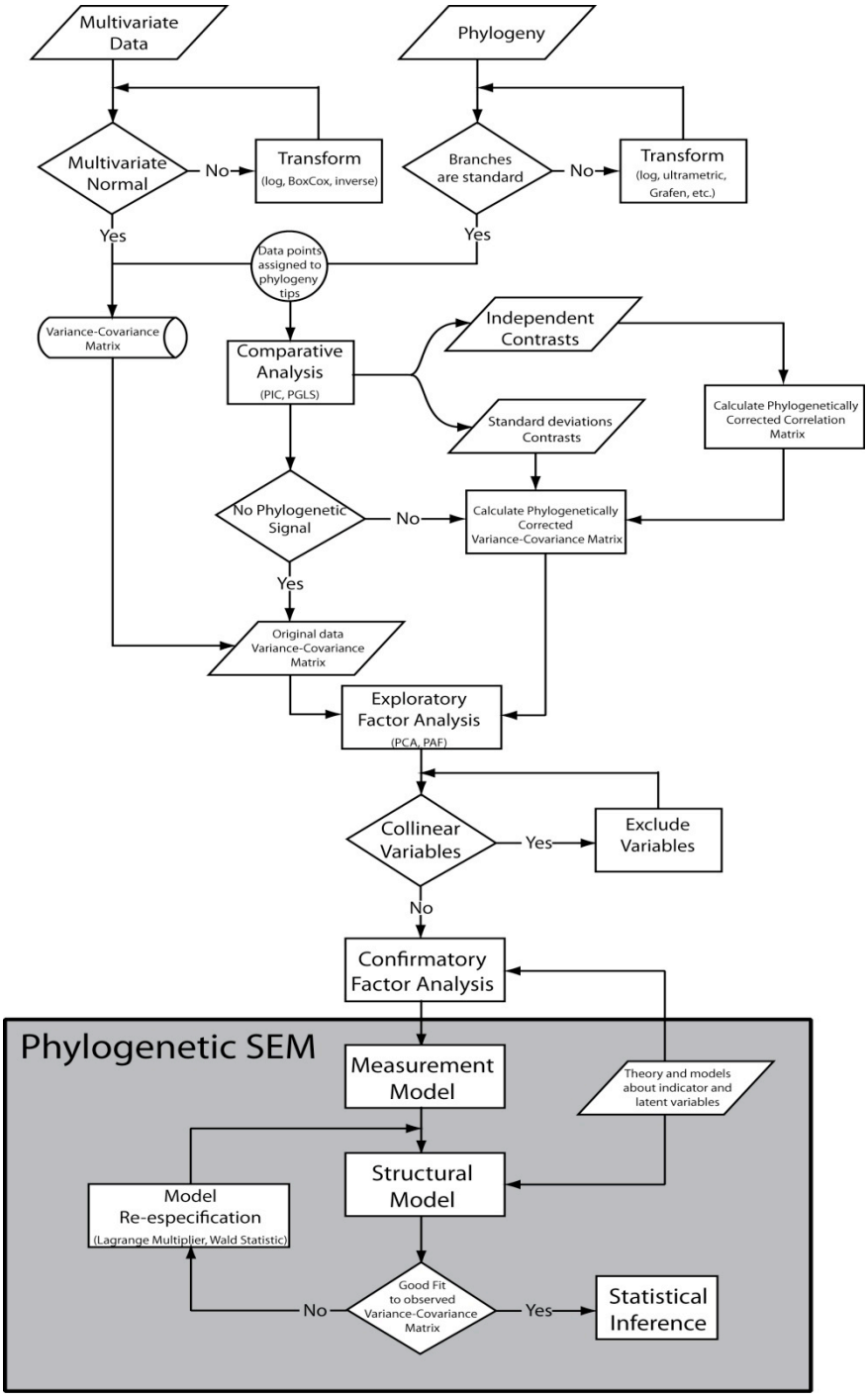

Figure 4: Flowchart of the steps necessary to perform a Phylogenetic SEM from raw multivariate data and phylogenetic reconstruction to factor analysis, confirmatory factor analysis, and SEM. Different steps are summarized from the different chapters developed in this report.

References

Adams, D. C. 2008. Phylogenetic meta-analysis. Evolution 62:567-572.

Armbruster, W. S., C. Pelabon, T. F. Hansen, and C. P. H. Mulder. 2004. Floral integration, modularity, and accuracy: distinguishing complex adaptations from genetic constraints. Pp. 23-50 *in* M. Pigliucci, and K. Preston, eds. Phenotypic Integration: Studying The Ecology. Oxford University Press, New York, USA.

Bentler, P. M. 1990. Comparative fit indexes in structural models. Psychological Bulletin 107:238-246.

Bentler, P. M., and D. G. Bonnett. 1980. Significance tests and goodness of fit in the analysis of covariance structures. Psychological Bulletin 88:588-606.

Björklund, M. 1997. Are "comparative methods" always necessary? Oikos 80:607-612.

Blomberg, S., T. Garland, and A. Ives. 2003. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. Evolution 57:717-745.

Brooks, D. R. 1996. Explanations of homoplasy at different levels of biological organization. Pp. 339 *in* M. Sanderson, J, and L. Hufford, eds. Homoplasy: The Recurrence of Similarity in Evolution. Academic Press, San Diego, CA.

Brown, T. 2006. Confirmatory Factor Analysis for Applied Research. The Guildford Press, New York, NY.

Butler, M. A., and A. A. King. 2004. Phylogenetic comparative analysis: a modeling approach for adaptive evolution. Am Nat 164:683-695.

Cheverud, J., M. Dow, and 1985a. An autocorrelation analysis of genetic variation due to lineal fission in social groups of rhesus macaques. American Journal of Physical Anthropology 67:113-122.

Cheverud, J. M., M. M. Dow, , W. Leutenegger, and 1985b. The quantitative assessment of phylogenetic constraints in comparative analyses: sexual dimorphism in body-weight among primates. . Evolution 39:1335-1351.

Clobert, J., T. Garland, and R. Barbault. 1998. The evolution of demographic tactics in lizards: a test of some hypotheses concerning life history evolution. J Exp Biol 11:329-364.

Cox, R. M., S. L. Skelly, and H. B. John-Alder. 2003. A comparative test of adaptive hypotheses for sexual size dimorphism in lizards. Evolution 57:1653-1669.

Desdevises, Y., P. Legendre, L. Azouzi, and S. Morand. 2003. Quantifying phylogenetically-structured environmental variation.. Evolution 57:2647-2652.

Felsenstein, J. 1985. Phylogenies and the comparative method. Am Nat 125:1-15.

Felsenstein, J. 2004. Inferring Phylogenies. Sinauer Associates, Sunderland, MA.

Freckleton, R. P., P. H. Harvey, and M. Pagel. 2002. Phylogenetic analysis and comparative data: a test and review of evidence. Am Nat 160:712-726.

Garland, T., and S. C. Adolph. 1991. Physiological differentiation of vertebrate populations. Annu. Rev. Ecol. Syst. 22.

Garland, T., A. F. Bennett, and E. L. Rezende. 2004. Phylogenetic approaches in comparative physiology. J Exp Biol 208:3015-3035.

Garland, T., A. W. Dickerman, C. M. Janis, and J. A. Jones. 1993. Phylogenetic analysis of covariance by computer simulation. Syst Biol 42:265-292.

Garland, T., P. H. Harvey, and A. R. Ives. 1992. Procedures for the analysis of comparative data using phylogenetically independent contrasts. Syst Biol 41:18-32.

Garland, T., and A. R. Ives. 2000. Using the past to predict the present: confidence intervals for regression equations in phylogenetic comparative methods. Am. Nat. 155:346-364.

Garland, T., P. E. Midford, and A. R. Ives. 1999. An introduction to phylogenetically based statistical methods, with a new method for confidence intervals on ancestral values. Amer. Zool. 39:374-388.

Gotelli, N. J., and A. M. Ellison. 2004. A Primer Of Ecological Statistics. Sinauer Associates, Sunderland, MA.

Grace, J. 2006. Structural Equation Modeling and Natural Systems. Cambridge University Press, Cambridge, UK.

Grafen, A. 1989. The phylogenetic regression. Philos. Trans. R. Soc. London 326:119-157.

Grant, P. 1999. Ecology and Evolution of Darwin's Finches. Princeton University Press, Princeton, NJ.

Hansen, T. F. 1997. Stabilizing selection and the comparative analysis of adaptation. Evolution 51:1341-1351.

Hansen, T. F., and E. P. Martins. 1996. Translating between microevolutionary process and macroevolutionary patterns: the correlation structure of interspecific data. . Evolution 50:1404-1417.

Hansen, T. F., J. Pienaar, and S. H. Orzack. 2008. A comparative method for studying adaptation to a randomly evolving environment. Evolution 62:1965-1977.

Harvey, P. H., and M. D. Pagel. 1991. The Comparative Method in Evolutionary Biology. Oxford University Press, Oxford.

Harvey, P. H., and A. Purvis. 1991. Comparative methods for explaining adaptations. Nature 351:619-624.

Hillis, D. M. 1998. Taxonomic sampling, phylogenetic accuracy, and investigator bias. Syst. Biol. 47:3-8.

Housworth, E. A., E. P. Martins, and M. Lynch. 2004. The phylogenetic mixed model. Am Nat 163:84-96.

Hu, L., and P. M. Bentler. 1999. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Structural Equation Modeling: A Multidisciplinary Journal 6:1-55.

Huxley, J., 1942, and 1942. Evolution: The Modern Synthesis. George Allen and Unwin, London, UK.

Isaac, N. J. B., S. T. Turvey, B. Collen, C. Waterman, and J. E. M. Baillie. 2007. Mammals on the EDGE: conservation priorities based on threat and phylogeny. . PLoS ONE 2:e296.

Jackson, J. E. 2003. A User's Guide to Principal Components. Wiley-Interscience, Hoboken, NJ.

Jolliffe, I. T. 2002. Principal Component Analysis, New York, NY.

Joreskog, K. G. 1993. Testing structural equation models. Pp. 294-316 *in* K. Bollen, and J. Lang, eds. Testing structural equation models. Sage, Newbury Park, CA.

Kline, P. 1994. An Easy Guide to Factor Analysis. Routledge Taylor & Francis Group, New York, NY.

Kline, R. 2005. Pinciples and Practice of Structural Equation Modeling. The Guilford Press, New York, NY.

Lajeunesse, M. J. 2009. Meta-analysis and the comparative phylogenetic method. Am Nat 174:369-381.

Li, C. 1975. Path Analysis: A Primer. Boxwood Press, Pacific Grove, CA.

Losos, J. B. 1999. Uncertainty in the reconstruction of ancestral character states and limitations on the use of phylogenetic comparative methods. Anim Behav 58:1319-1324.

Losos, J. B. 2000. Ecological character displacement and the study of adaptation. Proc Natl Acad Sci U S A 97:5693-5695.

Maddison, W. P. 1990. A method for testing the correlated evolution of two binary characters: are gains or losses concentrated on certain branches of a phylogenetic tree? Evolution 44:539-557.

Maddison, W. P., P. E. Midford, and S. P. Otto. 2007. Estimating a binary character's effect on speciation and extinction. Syst Biol:5.

Martins, E., and T. F. Hansen. 1997. Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. Am Nat 149:646-667.

Martins, E. P., , and T. Garland. 1991. Phylogenetic analyses of the correlated evolution of continuous characters: a simulation study. Evolution 45:534-557.

Martins, E. P., E. A. Housworth, and 2002. Phylogeny shape and the phylogenetic comparative method. Syst Biol 51:873-880.

Mayr, E. 1982. The Growth of Biological Thought: Diversity, Evolution, and Inheritance. Belknap Press of Harvard University Press, Cambridge, MA.

Mulaik, S. 2009. Foundations of Factor Analysis. Chapman & Hall/CRC.

Pagel, M. 1994. Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete data. Proc. R. Soc. London B 255:37-45.

Pagel, M. 1997. Inferring evolutionary processes from phylogenies. Zoologica Scr. 26:331-348.

Pagel, M. 1999. Inferring the historical patterns of biological evolution. Nature 401:877-884.

Pagel, M., and A. Meade. 2007. BayesTraits Manual. http://www.evolution.rdg.ac.uk/BayesTraits.html. School of Biological Sciences. University of Reading, Reading, UK.

Pagel, M., A. Meade, and D. Barker. 2004. Bayesian estimation of ancestral character states on phylogenies. Syst Biol 53:673-684.

Paradis, E., and J. Claude. 2002. Analysis of comparative data using generalized estimating equations. J Theor Biol 218:175-185.

Price, T. 1997. Correlated evolution and independent contrasts. Phil. Trans. R. Soc. London B 362:519-529.

Quinn, G. P., and M. J. Keough. 2002. Experimental Design and Data Analysis for Biologists. Cambridge University Press, Cambridge, UK.

Rao, D., C, M. Province, and A. 2000. The future of path analysis, segregation analysis, and combined models for genetic dissection of complex traits. Hum Hered 50:34-42.

Redding, D. W., and A. O. Mooers. 2006. Incorporating evolutionary measures into conservation prioritization. Conserv Biol 20:1670-1678.

Rohlf, F. J. 2001. Comparative methods for the analysis of continuous variables: geometric interpretations. Evolution 55:2143-2160.

Steel, M., A. Mimoto, and A. O. Mooers. 2007. Hedging one's bets: quantifying a taxon's expected contribution to future phylogenetic diversity. Evol Bioinform Online 3:237-244.

Steiger, J. H., and J. C. Lind. 1980. Statistically based tests for the number of common factors. Spring Meeting of the Psychometric Society, Iowa City, IA.

Tabachnick, B. G., and L. S. Fidell. 2007. Using Multivariate Statistics. Pearson Education, Inc., Boston, MA.

Tucker, L. R., and C. Lewis. 1973. The reliability coefficient for maximum likelihood factor analysis. Psychometrika 38:1-10.

Wake, D. B. 1991. Homoplasy: the result of natural selection, or evidence of design limitations? Am. Nat. 138:543-567.

Westoby, M., M. R. Leishman, and J. M. Lord. 1995a. Further remarks on phylogenetic correction. J Ecol 83:727-729.

Westoby, M., M. R. Leishman, and J. M. Lord. 1995b. Issues of interpretation following phylogenetic correction. . J Ecol 83:892–893.

Westoby, M., M. R. Leishman, and J. M. Lord. 1995c. On misinterpreting the "phylogenetic correction.". J Ecol 83:531–534.

Wright, S. 1918. On the nature of size factors. Genetics 3:367-374.

Wright, S. 1920. The relative importance of heredity and environment in determining the birth weight of guinea pigs. Proc Natl Acad Sci U S A 6:320-332.

Wright, S. 1921. Correlation and causation. Journal of Agricultural Research 10:557-585.

Wright, S. 1923. The theory of path coefficients: a reply to Niles' criticism. Genetics 8:239-255.

Wright, S. 1925. Corn and Hog Correlations. . Pp. 1-60. U.S. Department of Agriculture Bulletin, Washington, D.C.

Wright, S. 1934. The method of path coefficients. Ann Math Stat 5:161-215.

Wright, S. 1983. On "Path analysis in genetic epidemiology: a critique". Am J Hum Genet 35:757-768.

Vita

Juan Carlos Santos was born in Quito, Ecuador, on September 4th 1977. He is one of the three sons of Ernesto Santos and Sara García. He lived in his native city since his childhood. His love for living things was inspired by his visits with his father to the cloud forests in the Eastern Andean Foothills. His parents gave him, since early childhood, free access to books including Science and History. His love for basic sciences and the unselfish support from his parents influenced him to follow the path of Academia. He graduated from the San Gabriel High School in 1995 and he went for a full year visit to the US. In 1996, he studied Biology at the Pontificia Universidad Católica del Ecuador. He earned a B.A. in Biology in 2002 under the direction of Luis Coloma. In the fall of 2002 he entered the Program of Ecology, Evolution and Behavior at the University of Texas at Austin and got his Ph.D. in 2009 under the supervision of David Cannatella. He also took courses in statistics and he hoped to finish his M.S. before his Ph.D., but he did it a semester after. He met his wife Natalia Biani in Austin and together, they have been sharing their love for Nature and Science.

Permanent address: Ventura Aguilera N57-11 y Anonas, Quito, Pichincha, Ecuador.

This report was typed by the author.